О.Е. Елисеева, К.В. Русецкий (Минск, БГУИР)

ПОДХОДЫ К ФОРМАЛИЗАЦИИ ЕСТЕСТВЕННОГО ЯЗЫКА НА СЕМАНТИЧЕСКОМ ГРАФОВОМ ЯЗЫКЕ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

В настоящее время актуальными являются задачи машинной (компьютерной) обработки, анализа и синтеза текстов естественного языка. Для решения указанных задач разрабатываются различные формальные средства и языки представления знаний о закономерностях естественного языка (ЕЯ) и его конструкций: слов, словосочетаний, текстов, а также правил их образования. При этом значительные усилия прилагаются для того чтобы, с одной стороны, указанные формальные средства были просты и понятны человеку, а с другой — являлись хорошим базисом для компактного хранения и эффективной обработки знаний о ЕЯ в памяти компьютера. Немаловажное значение имеет также степень наглядности формального описания (представления) конструкций соответствующего языка. Одним из наиболее наглядных способов формализации информации и знаний является представление в виде семантической сети.

Представить любые знания в виде семантической сети позволяет формальный графовый семантический язык Semantic Code (SC) и его подъязык, являющийся соответствующим графическим (графовым) представлением, Semantic Code Graphical (SCg) [1]. Семантическая сеть, описанная с помощью языка SC, представляет собой графовую структуру, вершинами (узлами) которой являются знаки объектов предметной области, а дуги задают отношения между ними. Семантика отношений задается также соответствующим набором узлов с идентификаторами. Семантические сети языка SC являются однородными и имеют базовую теоретико-множественную интерпретацию, т.е. узлы интерпретируются как знаки множеств, а базовым отношением является отношение принадлежности множеству. Для описания семантики многих отношений допускается вхождение дуги в дугу. Такие вхождения дуг в дуги в терминах языка SC называются атрибутами.

Язык SC является универсальным базовым языком для создания на его основе множества различных подъязыков, позволяющих формализовывать знания различных предметных областей. Для создания такого подъязыка необходимо сформировать набор узлов с различными идентификаторами и типами семантических связей, которые будут однозначно описывать метаотношения соответствующей предметной области. Так, например, для задания множества всех глаголов того или иного языка достаточно ввести узел языка SC (sc-узел) с идентификатором «глагол». Любая дуга, выходящая из узла «глагол» и входящая в любой другой узел будет означать, что соответствующий узел является знаком некоторого определенного глагола. В линейном виде такое простейшее семантическое отношение (по сути, отношение типа «множество — элемент») представляется следующим образом:

глагол → бежать

В данной работе предлагается использовать язык SC для формального описания ЕЯ [2], а также естественно-языковых конструкций: предложений и текстов. Метаинформация о языке (например, морфологические характеристики и синтаксические роли слов) описывается по принципам, описанным выше примером.

Слово является основной номинативной единицей языка [3]. Следовательно, формализацию естественного языка разумно начать с формализации его (языка) лексического состава. Как указывалось в работе [2], начальной стадией формализации любой предметной области является рассмотрение ее базовой терминологии. Поэтому для начала уточним понимание используемых базовых понятий о лексическом составе языка.

Под **лексемой** будем понимать множество словоформ, объединенных общим лексическим значением. Под **словоформой** будем понимать множество реализаций некоторой парадигматической формы в предложениях языка. Реализацию словоформы в предложении будем называть **словом**. Каждой лексеме, словоформе и слову соответствует узел семантической сети — знак этой лексемы, словоформы или слова, соответственно. Чтобы различать в рамках семан-

тической сети слово, обозначающее некоторое понятие и знак этого понятия, которому также может быть сопоставлен узел семантической сети (см. пример выше), договоримся для идентификации узла словоформы или лексемы использовать соответствующие слова ЕЯ с префиксами f_ и l_ (от англ. form–форма и lexeme–лексема) соответственно. Узел, обозначающий конкретное слово, имеет совпадающее с этим словом текстовое содержимое. Это необходимо для реализации различных функциональных возможностей естественно-языковой системы. Пример формализованного описания лексемы с использованием графового языка SCg приведен на рисунке 1.

На рис. 1 видно, что под общим лексическим значением (лексемой) «брюки» объединяется 5 словоформ — реализаций данной лексемы в текстах русского языка. Так как формы именительного и винительного падежей для лексемы «брюки» совпадают, то фактически, так же, как в естественном языке мы имеем не 6, а 5 словоформ.

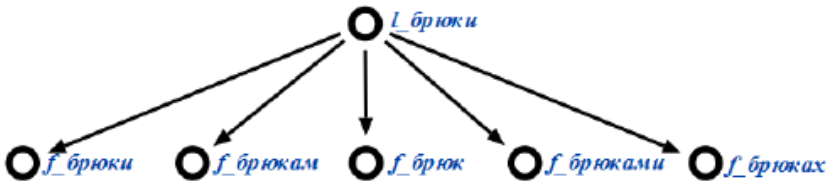

Рис. 1. **Фрагмент семантической сети, записанный на SCg и описывающий лексему «брюки».**

Известно, что все словоформы лексемы образуют парадигму, состав которой определяется набором изменяемых морфологических признаков лексемы (в случае существительного — это число и падеж). Парадигму и лексему, несмотря на кажущееся сходство, будем рассматривать как разные множества словоформ. Фрагмент парадигмы для лексемы «брюки» приведен на рис. 2.

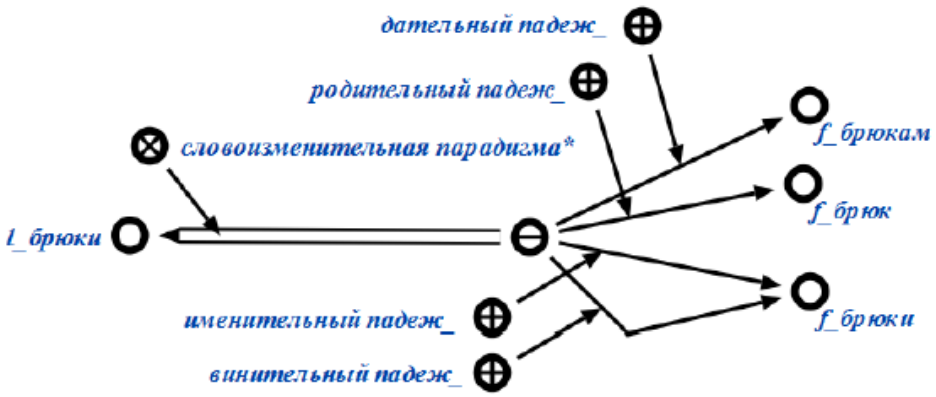

Рис. 2. **Фрагмент словоизменительной парадигмы слова «брюки».**

На рис. 2 сдвоенная дуга с атрибутом «словоизменительная парадигма*» обозначает связку асимметричного бинарного отношения (где связываемые элементы неравноправны). Морфологические признаки (на данном примере это наименования падежей) задаются атрибутами, которые назначаются дугам, связывающим парадигму с ее элементами. Обратим внимание на то, что форма

«брюки» имеет кратное вхождение в парадигму с кратностью, равной двум. Таким образом показано совпадение форм именительного и винительного падежа.

Аналогичным образом описываются различные семантические метаотношения между лексемами. Примерами таких отношений являются следующие: «являться синонимами», «являться омонимами», «являться гиперонимом». Например, на рисунке 3 представлено формальное описание на языке SC высказывания «Лексема “одежда» является гиперонимом для лексемы “брюки”«.

Рис. 3. **Пример формализации высказывания относительно двух лексем.**

Неизменяемые морфологические признаки формализуются на языке SC как принадлежность лексемы ко множеству лексем, имеющих некоторый общий неизменяемый морфологический признак. Например, на рис. 4 показано, что лексема «брюки» является элементом множества неодушевленных существительных.

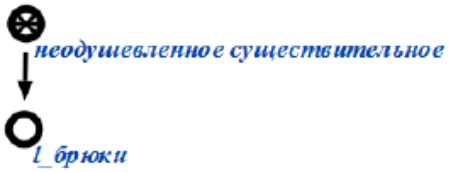

Рис. 4. **Пример формализации неизменяемого морфологического признака.**

Рассмотрим кратко особенности формального описания текстов естественного языка на языке SC.

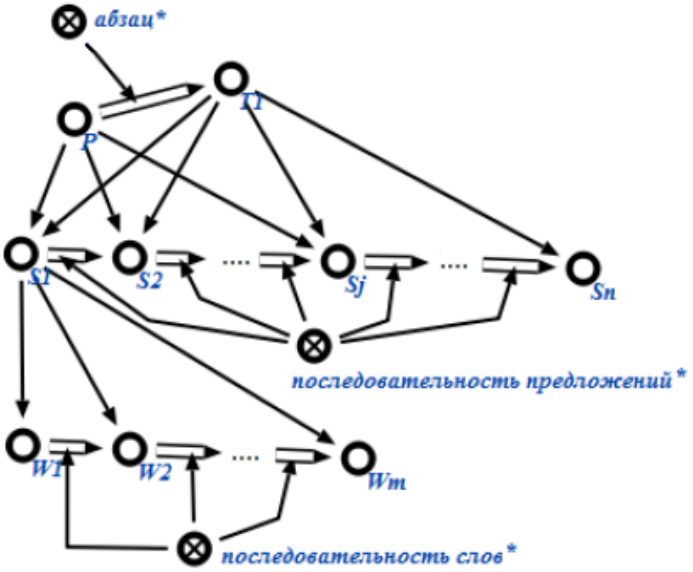

Рис. 5. **Пример формализации текста.**

Каждый текст T_i структурно представляет собой ориентированное множество предложений: $T_i = \langle S_1, S_2, ..., S_j, ..., S_n \rangle$. Предложение S_i в составе текста является последовательностью слов: $S_i = \langle W_1, W_2, ..., W_m \rangle$. Абзац P есть подмноже-

ство предложений текста, выделяемое на письме красной строкой: $P = \langle S_1, S_2, ..., S_j \rangle, P \subseteq T_1$.

Пример графового описания представленных выше утверждений представлен на рисунке 5. Информация о членах предложения и синтаксических отношениях представляется в виде соответствующих атрибутов дуг принадлежности, связывающих предложение и входящие в его состав слова.

Отметим также, что в рамках предлагаемого подхода разделители (пробелы, знаки препинания) включать в описание формальное предложений не требуется. Это связано, во-первых, с тем, что те части предложения, которые отделяют знаки препинания, можно (и нужно) формализовать как подмножества соответствующих структур. Так, например, простая часть сложносочиненного предложения является подмножеством предложения. Во-вторых, знак препинания в конце предложения определяется типом предложения — вопросительное, восклицательное либо повествовательное. Следовательно, достаточно лишь указать, что предложение является элементом множества предложений соответствующего типа. Таким образом, мы разрешаем проблему различного представления вопросительных и восклицательных предложений в разных языках (ярким примером здесь служит испанский язык, в котором в начале таких предложений ставится соответствующий перевернутый знак препинания).

В заключение отметим, что описанный выше подход едва ли является универсальным для всех языков мира. Он был апробирован для русского и английского языков, и есть предположения, что он работоспособен и для подобных им языков, однако его применимость к языкам вроде арабского пока не исследовалась.

ЛИТЕРАТУРА

1. Голенков, В.В. Представление и обработка знаний в графодинамических ассоциативных машинах. — Минск, 2001.

2. Елисеева, О.Е. Естественно-языковой интерфейс интеллектуальных систем : учеб. пособие / под науч. ред. проф. В.В. Голенкова. — Минск, 2009.

3. Розенталь, Д.Э. Современный русский язык. — 11-е изд. — М., 2009.